International Journal of Engineering, Sciences and Technologies
ISSN: 3079-2045
https://www.ij-est.org Volume 4, issue 3, https://doi.org/10.70961/WZUI0342

Prédiction de fraude de performance par machine learning des Centres de Santé de Base bénéficiaires du Financement Base sur la Performance

F. E. Rabearitsoa [1], A. Ratovohery [2], F. M. Randriatsarafara [3], M. A. Rakotomalala [4]

1. *RABEARITSOA Florent Eric, doctorant à l'Université d'Antananarivo, école doctorale STII, EAD/RIMSS, École Supérieure Polytechnique d'Antananarivo*
2. *RATOVOHERY Andry, doctorant à l'Université d'Antananarivo, école doctorale STII, EAD/RIMSS, École Supérieure Polytechnique d'Antananarivo.*
3. *RANDRIATSARAFARA Fidiniaina Mamy, professeur agrégé à la faculté de médecine de l'Université d'Antananarivo*
4. *RAKOTOMALALA Mamy Alain, professeur à l'École Supérieure Polytechnique d'Antananarivo, école doctorale STII, EAD/RIMSS*

Auteur correspondant : *RABEARITSOA Florent Eric,* Antananarivo (101), Madagascar
mail : eric.rabearitsoa2@gmail.com , WhatApp : +261341281751

RÉSUMÉ

L'intelligence artificielle (IA) a considérablement transformé le secteur de la santé, pourtant son utilisation dans le contexte du Financement basé sur la performance (FBP) est très peu documentée. Cette étude vise à utiliser l'IA pour rendre plus efficient le mécanisme de vérification des données, pour détecter de fraudes des Centres de santé de base (CSB) du FBP.

Méthodologie : (1) Évaluer la performance (régression logistique, XGBoost), combinée à la gestion du déséquilibre de classes par (SMOTE, et ROSE). (2) Évaluer à l'aide des métriques clés (AUC, F2-Score, et coût attendu) les six pipelines de classification de fraude ainsi construits. Leurs performances moyenne sont évaluées par les tests de Friedman, de Nemenyi, et de DeLong, afin d'identifier le modèle de classification de fraudes des CSB le plus robuste, et performant. (3) Vérifier par le test du Chi-deux le lien entre l'anomalie des données, et la fraude.

Les résultats montrent que : le lien entre fraude, et anomalie des données est significatif. Et le modèle XGBoost est classé plus performant [accurancy (96%), balanced accuracy (91%), sensibilité (86%), spécificité (96%), F2-score (84%), l'AUC (95%), et Kappa de Cohen (78%)]. L'utilisation de l'IA permet de réduire à moins 25% le nombre de CSB à vérifier du FBP.

L'anomalie de données est un indice de suspicion de fraude, et le modèle XGBoost est classé performante et stable pour détecter les fraudes des CSB. Cette étude contribue à optimiser l'allocation des ressources publiques affectées au FBP par l'utilisation de l'IA.

Mots-clés

Financement basé sur la performance, achat stratégique, régression logistique, XGBoost, économie de la santé

ABSTRACT

Artificial Intelligence (AI) has significantly transformed the healthcare sector; however, its use within the context of Performance-Based Financing (PBF) remains poorly documented. This study aims to use AI to improve the efficiency of data verification mechanisms and fraud detection among Primary Health Centers (PHC) involved in PBF.

Methodology: (1) Evaluate model performance using logistic regression and XGBoost, combined with class imbalance management techniques (SMOTE and ROSE). (2) Assess six fraud classification pipelines using key metrics (AUC, F2-score, and expected cost). Their average performance is further evaluated using Friedman, Nemenyi, and DeLong statistical tests to identify the most robust and best-performing fraud classification model. (3) Verify the relationship between data anomalies and fraud using the Chi-square test.

Results: The findings show that the relationship between fraud and data anomalies is statistically significant. The XGBoost model is ranked as the best-performing model [accuracy (96%), balanced accuracy (91%), sensitivity (86%), specificity (96%), F2-score (84%), AUC (95%), and Cohen's Kappa (84%)]. The use of AI reduces by at least 25% the number of PHC facilities requiring verification under PBF.

Data anomalies remain an indicator of suspected fraud in PHC, and XGBoost demonstrates superior and stable fraud classification performance. This study contributes to optimizing the allocation of public resources assigned to PBF through the use of AI.

Keywords

Performance-based financing (PBF), Strategic health purchasing, Logistic regression, Extreme Gradient Boosting (XGBoost), Health Economics

I. INTRODUCTION

À Madagascar depuis 2013, plusieurs Centres de santé de base (CSB) sont sous contrat de performance du Financement Basé sur la Performance. Ce mécanisme repose sur l'achat stratégique de performance réelle des formations sanitaires [1]. L'achat de performance des indicateurs sanitaires des CSB repose sur la vérification, ou audit des données sanitaires déclarées. Ce mécanisme de verification systématique garantit la fiabilité des indicateurs sanitaires déclarés objet d'achat. Ce mécanisme limite les risques de comportements frauduleux des prestataires, car il implique également des enquêtes auprès des usagers, afin de confirmer la véracité de la prestation de soins offerte. Ce mécanisme est plus sûr pour lutter contre les fraudes des CSB [2]. Pourtant ce mécanisme nécessite une mobilisation de ressources importantes. La fraude de performance du FBP consiste à créer des usagers fictifs, ou falsifier les données sanitaires afin d'augmenter les primes du personnel.

L'avancée majeure de l'intelligence artificielle (IA) a permis de développer considérablement la lutte contre les fraudes dans différents domaines. Et la détection de façon précoce de certaines anomalies des données est une piste d'orienter à la détection de risque de cas de fraudes [3,4].

L'utilisation de l'apprentissage supervisé dans le domaine du Financement Basé sur la Performance (FBP) du secteur sanitaire est peu documenté, cependant son utilisation est une piste pour réduire le coût de la vérification externe des données des CSB, en limitant l'audit ou la verification des données aux CSB à risque de fraude.

Cette étude a évalué la performance de deux modèles d'apprentissage supervisé, *la régression logistique*, et le *XGBoost,* en combinant à la gestion de déséquilibre de classes par la méthode de SMOTE (Synthetic Minority Over-sampling Technique), et de ROSE (Random Over Sampling Examples), plus la détection d'anomalies des données par l'algorithme « *Local Outlier Factor* » afin de déterminer un modèle de classification de fraude des CSB, adapté au contexte FBP à Madagascar.

La présente étude explore l'utilisation de modèles d'apprentissage supervisé afin de réduire le coût de verification classique des CSB en utilisant l'IA pour cibler les CSB à risque de fraude. Et restreindre la verification aux Centres présentant une probabilité élevée de risque de fraudes par création des usagers fictifs.

Cette étude se positionne à la croisée de l'économie de la santé, et de l'intelligence artificielle appliquée, en proposant une méthode novatrice visant à optimiser l'efficience d'allocation de ressources affectées au FBP à Madagascar.

II. RAPPEL SUR L'ACHAT DE PERFORMANCE DES CSB

Trimestriellement, une verification des données sanitaires, suivi d'une évaluation de qualité de service, puis d'une enquête auprès des usagers ont été réalisées un niveau de chaque CSB.

a) Verification des données

C'est la verification des indicateurs sanitaires, ou audit des données qui est réalisée systématiquement par une équipe externe pour auditer les données (indicateurs de santé) des CSB avant l'achat de performance. Chaque audit des données a été suivi d'une évaluation de la qualité de services de chaque CSB, et une enquête auprès des usagers.

b) Enquête auprès des usagers de CSB

L'enquête auprès des usagers a pour objectifs de vérifier la véracité de prestation de soins enregistrés dans les CSB, et évaluer la satisfaction des usagers. Pour la vérification de prestation, elle a permis de classer les CSB en : (1) **CSB frauduleux** par création des usagers fictifs, (2) **CSB fortement suspects de fraudes**, et (3) ***CSB conformes.***

Le schéma ci-dessous illustre le cycle trimestriel d'achat de performance des CSB.

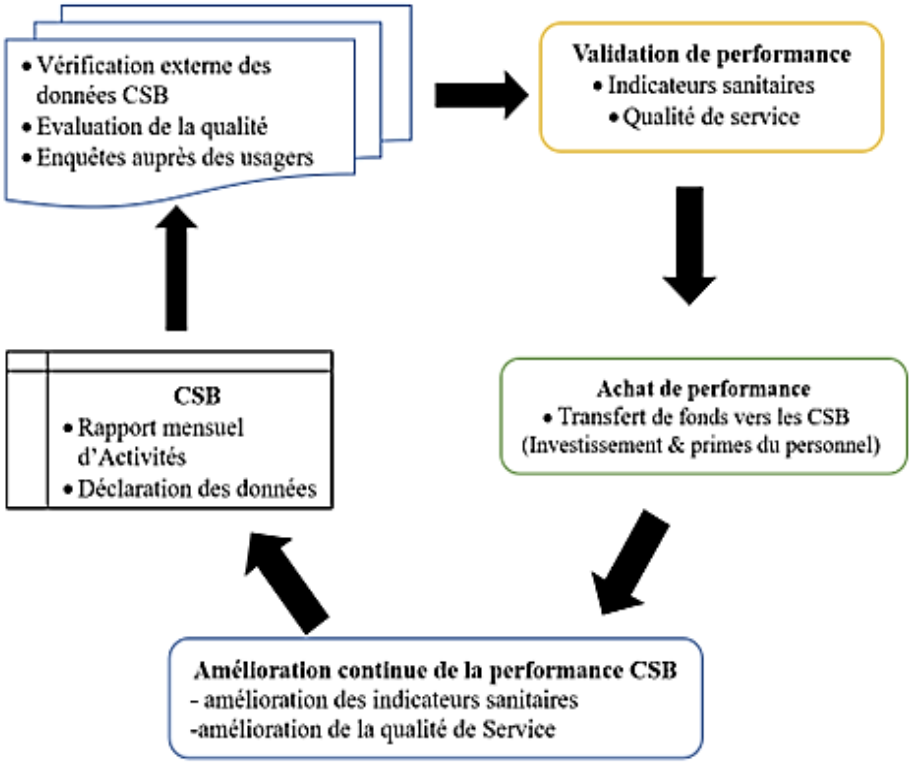

Figure N° 1 : Cycle trimestriel d'achat de performance des CSB

III. OBJECTIFS DE L'ÉTUDE

a) L'objectif général consiste à réduire le coût de la verification classique des CSB enroulés dans l'approche du FBP.

b) Les objectifs spécifiques consistent à :

1. identifier un modèle de classification de fraude des CSB basé sur l'apprentissage supervisé adapté au contexte du FBP ;
2. réduire à 25% le nombre de CSB à vérifier en utilisant l'IA, afin de rendre plus efficient le mécanisme de verification externe des données.

IV. REVUE DE LITTÉRATURE

a) Local Outlier Factor (LOF)

En 2000, Breunig et al [5] ont proposé pour la première fois la notion Local Outlier Factor (LOF) dans le but de détecter une valeur aberrante locale (local outlier) qui est significativement moins dense que son voisinage immédiat.

Définition

Selon (Breunig et al, 2000) [5], *« le facteur d'isolement local (LOF) d'un objet p est défini comme la moyenne des rapports entre la densité de proximité locale des k plus proches voisins de p et la densité de proximité locale de p »*.

$$LOF_k(p) = \frac{1}{|N_k(p)|} \sum_{O \in N_k(p)} \frac{LRD_k(O)}{LRD_k(p)}$$

où :

- $N_k(p)$: l'ensemble des k-plus proches voisins de p,
- $\text{Reach} - \text{dist}_k(p, O) = max\{k - distance(O), d(p, O\}$: distance atteignable de p à O,
- $d(p,O)$: la distance euclidienne (ou autre métrique) entre p et du point O
- $LRD_k(0)$: Local Reachability Density du point 0 (densité locale atteignable du point 0)
- $LRD_k(p) = \left(\frac{\sum_{0 \in N_k(p)} \text{Reach-dist}_k(p,O)}{N_k(p)}\right)^{-1}$, c'est la densité locale atteignable (Local Reachability Density) de p

Interprétation de score LOF

- si $LOF_k(p) \approx 1$, la densité locale de p est similaire à celle de ses voisins → ***point normal***,
- si $LOF_k(p) > 1$, alors p est dans une région moins dense que ses voisins → ***outlier***,
- Plus la valeur LOF est élevée, plus le ***degré d'anomalie est élevé***.

b) Modèle de la régression logistique

Définition

La régression logistique est une méthode d'apprentissage supervisé permettant de modéliser la probabilité qu'une observation appartiennent à une classe. Elle applique une fonction sigmoïde à une combinaison linéaire des prédicteurs [6] :

$$P(y = 1|x) = \frac{1}{1 + e^{-x^{\top}\beta}}$$

où :

- $x \in \mathbb{R}^p$: vecteur des variables explicatives ou « features »
- $\beta \in \mathbb{R}^p$: vecteur des coefficients à apprendre du modèle ou « poids »
- $x^\top \beta = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p$; c'est le produit scalaire des entrées (combinaison linéaire)
- e : base des logarithmes naturels

Ce modèle mathématique utilisé en régression logistique comme fonction de prédiction permet de transformer une combinaison linéaire des variables en une probabilité entre [0 et 1], ou classification binaire des problèmes [7].

c) Modèle d'extreme gradient boosting

Définition

Selon l'étude réalisée par (Chen et al, 2016) [8] l'Extreme Gradient Boosting (XGBoost) est un algorithme d'apprentissage supervisé fondé sur le gradient boosting, conçu pour être hautement performant, efficace et scalable. Il optimise une fonction de perte différentiable par descente de gradient en espace fonctionnel, en combinant séquentiellement des arbres de décision régularisés. Le modèle introduit des mécanismes avancés tels que la régularisation explicite (L1/L2), l'élagage automatique de la taille des arbres de décision, la gestion native des valeurs manquantes et une implémentation parallèle optimisée, le rendant adapté aux très grands jeux de données. Le XGBoost utilise le modèle mathématique de type additif.

Soit un jeu de données d'apprentissage ci-dessous :

- $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$, avec $x_i \in \mathbb{R}^d$, y_i est une variable cible (binaire, continue ou mutilasse).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

où :

- F est l'ensemble des fonctions représentables par des arbres de la « Classification And Regression Trees » ou CART,
- $f_k \in F$, *et* f_k est un arbre de décision (fonction de *partition* de l'espace),
- K est le nombre total d'arbres (boosting rounds),
- $\hat{y}_i$ est la prédiction du modèle pour l'observation x_i.

V. MÉTHODOLOGIE

Premièrement : Pour les données d'enquêtes auprès des usagers, le test statistique de CHI-2 est utilisé pour déterminer l'existence éventuelle de lien entre la fraude et les scores LOF de la fréquentation globale de chaque CSB.

Deuxièmement : Nous avons utilisé deux modèles d'apprentissage supervisé (régression logistique et XGBoost), combinés à la gestion de déséquilibres des classes par la technique ROSE, et de SMOTE, afin de créer six pipelines de classification de fraude.

Liste de pipeline à évaluer :

- *Régression logistique sans rééquilibrage des classes (RL_none);*
- *Regression logistique avec rééquilibrage de classes par ROSE (RL_rose) ;*
- *Regression logistique avec rééquilibrage de classes par SMOTE (RL_smote) ;*
- *XGBoost sans rééquilibrage des classes (XGBoost_none) ;*
- *XGBoost avec rééquilibrage de classes par ROSE (XGBoost_rose) ;*
- *XGBoost avec rééquilibrage de classes par SMOTE (*XGBoost_smote).

- Traitement des données
- Division des données (80% entrainement -20% test)
- Entraînement, optimisation, et Évaluation de performance des six pipelines selon les métriques clés (AUC, F2-score, Coût attendu).
- Comparaison de performance moyenne afin de déterminer le modèle de classification de fraude adapté au contexte FBP par (le test de Friedman, test post-hoc de Nemenyi), selon les métriques clés d'évaluation. Et finalement, nous avons utilisé le test de Delong pour identifier le modèle plus performant selon l'AUC, parmi les modèles classés premiers selon les métriques clés.
- Une fois le modèle plus stable identifié, il sera validé par des données externes des sources de données d'entrainement et de test, en utilisant les données sanitaires des 22 CSB du District sanitaire d'Antanifotsy.

Source des données

- Données de 25 CSB enroulés dans l'approche FBP du District sanitaire d'Ambatolampy de 2019 à 2021 :
  - données des 11 enquêtes trimestrielles auprès des usagers (échantillonnage aléatoire stratifié, dont la taille d'échantillon est de 5% de la fréquentation globale du CSB de la période vérifiée) ;
  - données vérifiées ou auditées des CSB par des équipes externes ;
  - Scores d'évaluation de la qualité de service des CSB ;

- Données issues du système national d'information sanitaire (2019-2021).

Les sources de données du FBP sont des données primaires non publiées de l'approche FBP à Madagascar. Ces données internes n'ont, jusqu'à présent, fait l'objet d'aucune exploitation à des fins de modélisation par intelligence artificielle pour ciblage des CSB suspects de fraude du FBP.

Variables explicatives

- proportion de fréquentation mensuelle du CSB [fréquentation/ (population annuelle/12)] ;
- score LOF moyen, k de voisinage plus proches : $5 \leq k \leq 12$;
- proportion de fréquentation moyenne mensuelle du District ;
- score de qualité de Service du CSB ;
- activité de soins de proximité ;
- population du secteur sanitaire du CSB ;
- distance entre le CSB et le District ;
- nombre de Fokontany ;
- nombre de villages ;
- classement d'habitation de la population majoritaire selon la distance par rapport aux CSB (1 à 3) ;
- densité de la population par Village ;
- Trimestre d'évaluation (1 à 4) ;
- mois d'évaluation (1 à 12).

Résultats attendus

❖ **Résultats d'enquête auprès des usagers :**

- prévalence de fraude ;
- et lien entre fraude, et les scores LOF moyen de fréquentation des CSB

❖ **Test de Friedman et Nemenyi :**

Hypothèse du test de Friedeman

- *Hypothèse nulle (H_0) : il n'y a pas de différence significative entre les modèles ($p<5\%$) ;*
- *Hypothèse alternative (H_1) : Il existe au moins une différence significative entre les modèles ($p\geq5\%$).*

Hypothèse du test de Nemenyi

- *Hypothèse nulle (H_0) : Il n'y a pas de différence significative entre les deux modèles comparés ($p<5\%$) ;*

- *Hypothèse alternative (H_1) : Il existe au moins une différence significative entre les deux modèles comparés (p≥5%).*

Résultats attendu (Friedman et Nemenyi) :

- tableau final de performances moyenne selon les métriques clés d'évaluation (AUC, F2-Score, Coût attendu) ;
- classement de performance moyenne des pipelines, et CD (Critical Difference) de Nemenyi ;
- choix de pipeline plus performant selon les métriques clés (AUC, F2-Score, Coût attendu).

❖ **Test de DeLong**

Hypothèse du test de Delong

- *Hypothèse nulle (H_0) : il n'y pas de différence significative entre les AUC de deux modèles comparés (p < 5%) ;*
- *Hypothèse alternative (H_1) : il existe une différence significative entre les deux AUC de deux modèles comparés (p ≥ 5%).*

Résultats attendus :

- identification du pipeline final plus performant à utiliser pour la vérification ciblée des CSB par l'intelligence artificielle selon la métrique AUC.
- calculer les principales métriques de performance du modèle de classification des fraudes classé plus stable, et performant après le test de Delong : (Accuracy, Balanced Accuracy, Kappa de Cohen, NPV, PPV, Sensibilité, Spécificité, F2-Score, AUC, Cout attendu).

❖ **Validation par des données externes du modèle**

Confrontation des résultats d'enquête sur les fraudes des 22 CSB du District d'Antanifotsy du premier trimestre 2024, et les résultats de détection de fraude des CSB, utilisant le modèle de classification de fraude classé le plus performant après les différents tests statistiques.

Logiciels utilisés :

- RStudio version :2026.01.1.403, R version : 4.5.2
- Excel pour le stockage des données (csv)

VI. Résultats

a) Résultats d'enquêtes auprès des usagers (2019-2021)

1. Prévalence de fraude des CSB

Tableau 1 : Prévalence de fraude des CSB du District d'Ambatolampy (2019-2022)

Enquête CSB	Proportion
CSB Frauduleux	14%
CSB Conformes	86%

La prévalence moyenne pour la période de 2019 à 2021 est de 14%. Pour chaque enquête trimestrielle, elle a varié de 0% à 20%, soit moins de 25% des CSB du District d'Ambatolampy.

2. *Corrélation fraude des CSB et les scores LOF moyens*

Tableau 2 : Corrélation entre fraude des CSB et les scores LOF de fréquentation des CSB

FRAUDE				
LOF moyens	NON	OUI	fraude	p chi-2
< 1,1	456	62	12%	p = 0,029
≥ 1,1	225	49	14%	

k voisins plus proches (5<minPts<12), seuil des scores LOF moyen (1,1), test Chi-deux α=0,05

b) Test de Friedman

Tableau 3 : Résultats du test de Friedman selon les trois métriques clés sur les six pipelines

Métrique clé	p-value
F2-Score	p < 2,2 10-16
AUC	p < 2,2 10-16
Coût attendu	p< 2,2 10-16

c) Test de Nemenyi

❖ **Selon la métrique d'évaluation (F2-Score)**

Tableau 4 : Classement par rang moyen (F2-score)

Classement	Pipeline sur l'ensemble des blocs de CV	Rang moyen
1	Régression logistique (ROSE)	1,58
2	Régression Logistique (SMOTE)	2,13
3	XGBoost (ROSE)	3,21
4	Régression Logistique (none)	3,66
5	XGBoost (SMOTE)	4,72
6	Xgboost (none)	5,7

Tableau 5 : Matrice des comparaisons multiples (F2-score)

Métrique : F2-Score	RL_none	RL_rose	RL_smote	XGB_none	XGB_rose
RL_rose	$4{,}1\ 10^{-07}$	-	-	-	-
RL_smote	$6{,}2\ 10^{-4}$	0,68358	-	-	-
XGBoost_none	$7{,}4\ 10^{-07}$	$4{,}1\ 10^{-14}$	$4{,}8\ 10^{-14}$	-	-
XGBoost_rose	0,83578	$1{,}9\ 10^{-4}$	0,04505	$4{,}3\ 10^{-10}$	-
XGBoost_smote	0,05241	$6{,}3\ 10^{-14}$	$6{,}7\ 10^{-11}$	0,09261	$7{,}7\ 10^{-4}$

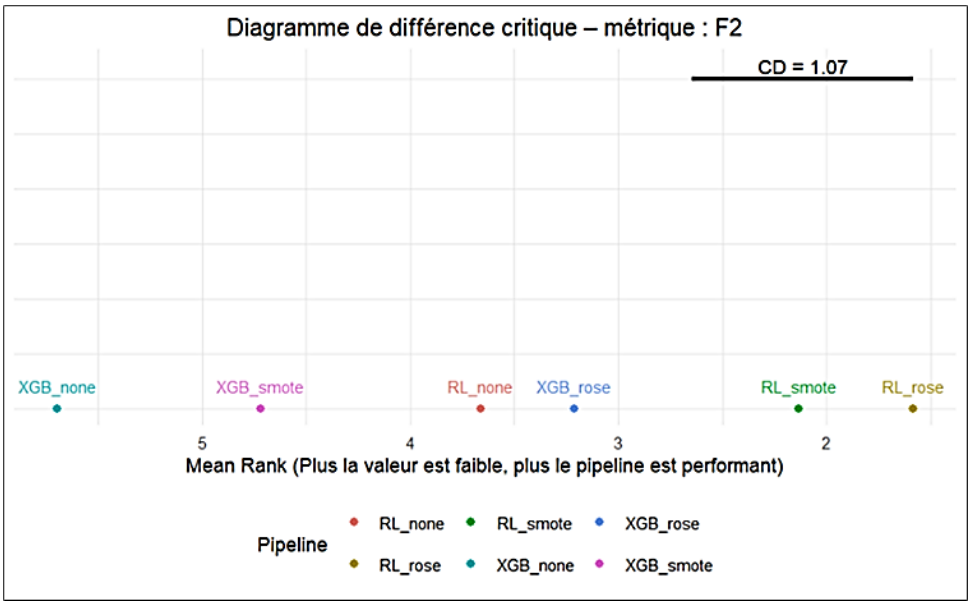

Figure 1 : Diagramme de différence critique (F2-score)

❖ **Selon la métrique d'évaluation (AUC)**

Tableau 6 : Classement par rang moyen (AUC)

Classement	Pipeline sur l'ensemble des blocs de CV	Rang moyen
1	Régression Logistique (ROSE)	2,01
2	Régression Logistique (SMOTE)	2,26
3	Régression Logistique (NONE)	2,6
4	Xgboost (ROSE)	3,95
5	Xgboost (SMOTE)	4,63
6	Xgboost (NONE)	5,55

Tableau 7 : Matrice des comparaisons multiples (AUC)

Métrique : AUC	RL_none	RL_rose	RL_smote	XGB_none	XGB_rose
RL_rose	0,61393				
RL_smote	0,94444	0,98539			
XGBoost_none	$1,3\ 10^{-13}$	$5,3\ 10^{-14}$	$4,6\ 10^{-14}$		
XGBoost _rose	$4,18\ 10^{-3}$	$3,2\ 10^{-06}$	$9,2\ 10^{-05}$	$2,7\ 10^{-4}$	
XGBoost_smote	$8,6\ 10^{-07}$	$3,8\ 10^{-11}$	$3,6\ 10^{-09}$	0,13654	0,45440

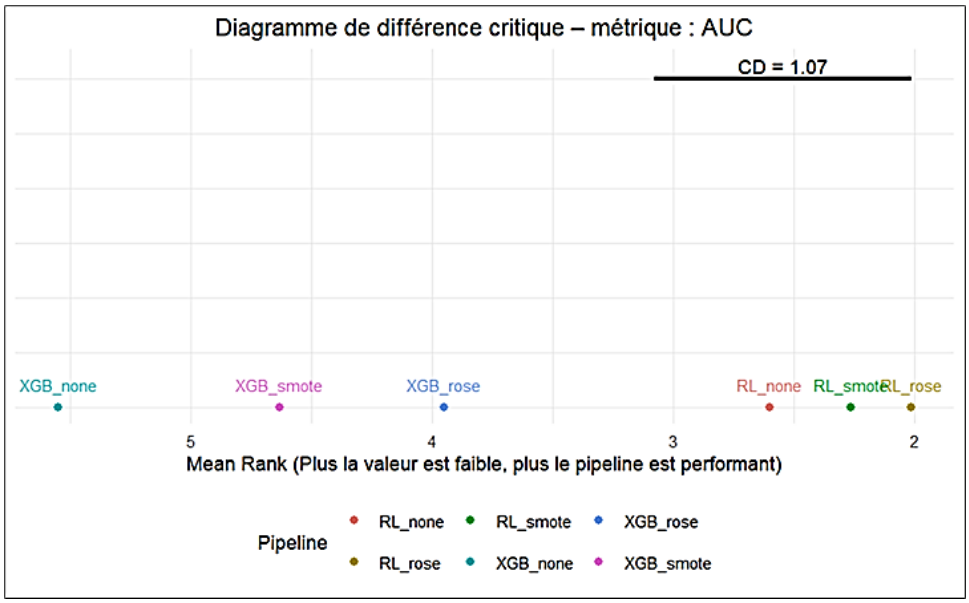

Figure 2 : Diagramme de différence critique (AUC)

❖ **Selon la métrique d'évaluation (Cout attendu)**

Tableau 8 : Classement par rang moyen (Coût attendu)

Classement	pipeline sur l'ensemble des blocs de CV	rang moyen
1	XGBoost _none	1,54
2	XGBoost _smote	2,46
3	RL_none	3,33
4	XGBoost _rose	3,65
5	RL_smote	4,76

6	RL_rose	5,26

Tableau 9 : Matrice des comparaisons multiples (Coût attendu)

Métrique : coût attendu	RL_none	RL_rose	RL_smote	XGB_none	XGB_rose
RL_rose	$3{,}7\ 10^{-06}$	-	-	-	-
RL_smote	$1{,}8\ 10^{-3}$	0,76477	-	-	-
XGBoost_none	$2{,}5\ 10^{-05}$	$5{,}1\ 10^{-14}$	$5{,}3\ 10^{-14}$	-	-
XGBoost_rose	0,95688	$2{,}4\ 10^{-4}$	0,03566	$2{,}6\ 10^{-07}$	-
XGBoost_smote	0,18383	$1{,}1\ 10^{-12}$	$1{,}2\ 10^{-08}$	0,13654	$1{,}8\ 10^{-2}$

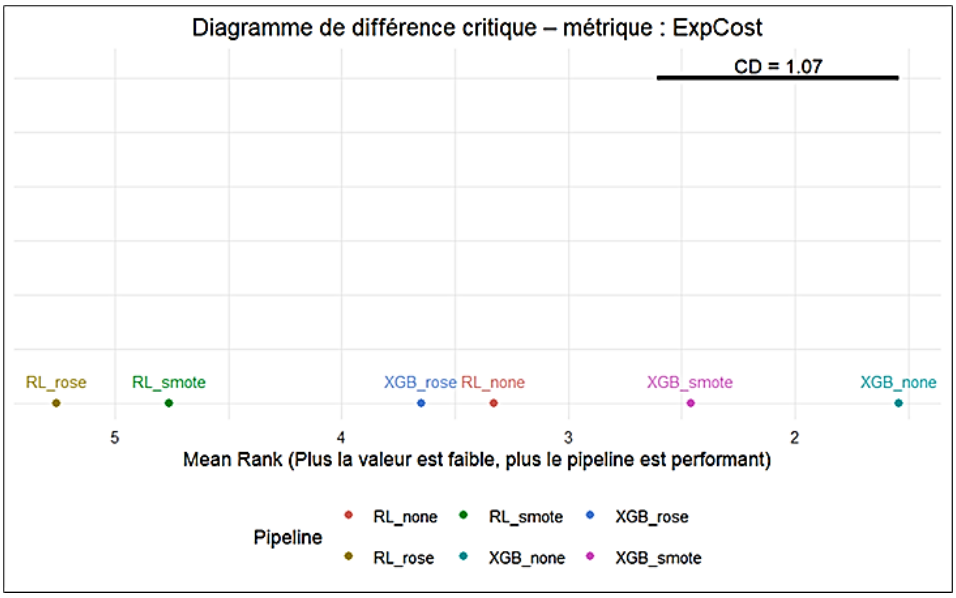

Figure 3 : Diagramme de différence critique (Coût attendu)

d) Test *de DeLong*

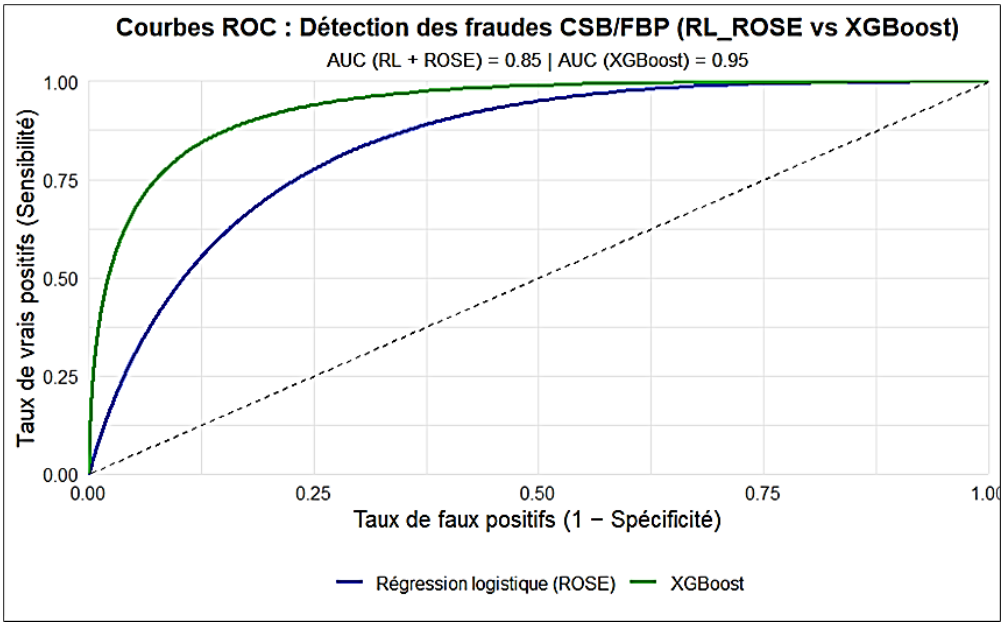

Figure 4 : Courbes ROC : régression logistique (ROSE) et XGBoost (None)

- Z = -2,205 (RL_rose, Xgboost_None), Z (négatif) signifie AUC RL_rose < AUC Xgboost
- p-value = 0,02745 (différence significative de AUR (85% vs 95%)
- Seuil de significativité α = 0,05

e) **Comparaison de performance moyenne de deux modèles de classification**

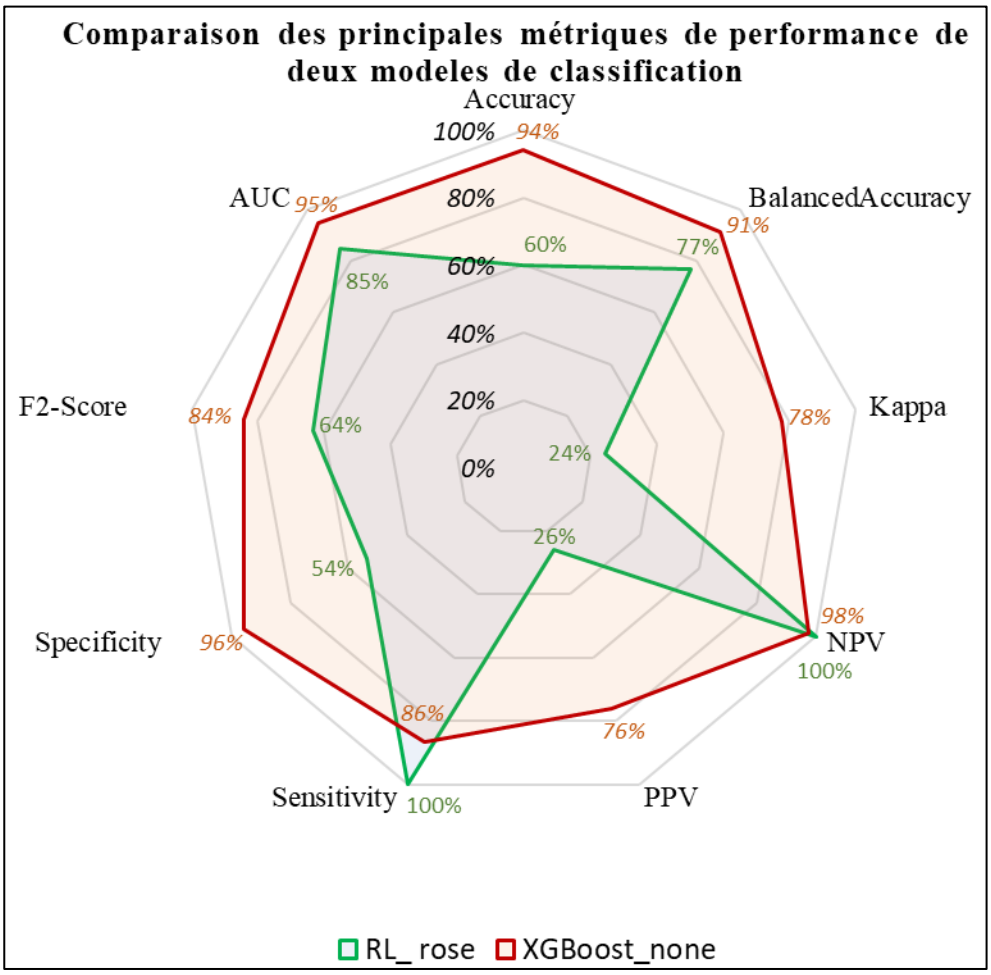

Figure 5 : Comparaison de performance de régression logistique (ROSE) et XGBoost (sans rééquilibrage de classes)

f) Validation externe du modèle XGBoost par des données du District sanitaire d'Antanifotsy

Au total 22 CSB sont sous contrat dans le District d'Antanifotsy. Le nombre de Centres de santé de base à vérifier par l'IA est limité à 5 CSB.

Données de référence sur les fraudes des CSB :

- Résultats d'enquête trimestrielle auprès des usagers de 22 CSB du District d'Antanifotsy ;
- Taille d'échantillon (1133 usagers enregistrés dans les CSB), date d'enquête (Novembre 2024), période vérifiée (T1/2024).

Tableau 10 : Résultats de validation externe

CSB	Frauduleux	Suspects
XGBoost	2	3
Résultats d'enquête auprès des usagers	2	5

VII. DISCUSSION

a) Rééquilibrage des données d'entrainement

Les contraintes liées au nombre d'observation des données utilisée nous ont motivé à opter le choix de sur-échantillonnage de classe minoritaire (fraude 14%), lors de rééquilibrages des données. Ce choix présente des avantages et des risques selon la technique de rééquilibrage utilisée. Les conséquences les plus fréquentes sont la perte d'information, l'instabilité du modèle, et le biais de distribution. Pourtant le rééquilibrage présente des principaux avantages tels que : réduit le biais vers la classe majoritaire, et l'amélioration de détection des classes minoritaires.

b) Test de Friedman

Les six pipelines ont été comparés à l'aide du test non paramétrique de Friedman sur les 50 blocs de validation croisée (10×5 = 50 blocs, CV folds), puis ils sont classés par le classement moyen ou (mean rank). Ce type de classement indique la performance relative et la stabilité des pipelines.

Les résultats du test de Friedman sur les six pipelines testés, selon les trois métriques clés d'évaluation (F2-Score, AUC, coût attendu) ont montré que pour chaque métrique d'évaluation, le test de Friedman est statistiquement significatif. La valeur de ($p < 2{,}2\ 10\text{-}16$), *Hypothèse nulle* (H_0) est rejetée, cela signifie qu'il existe une différence statistiquement significative de performance moyenne de six pipelines de classification de fraude des CSB.

c) Test de Nemenyi

Après le test de Friedman significatif, les résultats du test de Nemenyi nous permettent de déterminer le pipeline performant pour chaque métrique clé étudiée.

❖ ***Classement moyen selon la métrique « F2-Score »***

Pour la métrique d'évaluation F2-Score, le pipeline utilisant le modèle de régression logistique (ROSE) est classé premier, et le pipeline du modèle de régression logistique (SMOTE) en deuxième. Le résultat du test de Nemenyi indique que l'hypothèse nulle H_0 est confirmée. Le résultat du test post-hoc de Nemenyi confirme qu'aucune différence significative n'est détectée. Le classement observé de deux premiers pipelines peut être dû à la variabilité aléatoire des données. Par rapport aux

quatre pipelines restants, la supériorité en termes de classement moyen de deux premiers pipelines est statistiquement significative (*p-value* < 5%).

❖ ***Classement moyen selon la métrique « AUC »***

les modèles de régression logistique figurent en trois premières positions (ROSE en premier, SMOTE en deuxième, et sans rééquilibrage en troisième). Les résultats du test de Nemenyi montrent que la valeur p-value est supérieure à 5% pour chaque paire de pipeline. L'*hypothèse nulle* (H_0) est rejetée, tous les trois pipelines du modèle de régression logique ont une performance équivalente pour la métrique AUC. Les différences apparentes dans le classement peuvent être dues à la variabilité aléatoire des données. Cette supériorité de classement moyen de trois pipelines du modèle de régression logistique est statistiquement significative (*p-value* < 5%), par rapport aux trois pipelines du modèle Xgboost.

❖ ***Classement moyen selon la métrique « coût attendu »***

L'utilisation du machine learning pour cibler les CSB à risque de fraude de performance du FBP présente une particularité en termes de coût d'audit et de fraude. Dans le cas du FBP du secteur sanitaire, le coût d'une fraude non détectée (c'est le coût non récupéré de sanction pécuniaire appliquée au CSB frauduleux) étant largement supérieur à celui du coût d'un audit des données inutile déclenché par une fausse alerte. L'optimisation du seuil de décision a été guidée par la minimisation du coût attendu, avec une pénalisation fortement asymétrique des faux négatifs (FN), par rapport aux faux positifs (FP). Le paramètre d'optimisation utilisé est (coût attendu = 10*FN + 1*FP). Cette approche conduit logiquement à des seuils garantissant un taux de faux négatifs le plus faible possible, tout en conservant une capacité discriminante de l'AUC, et une performance du F2-score.

Les deux pipelines du modèle Xgboost occupent les premières positions (sans rééquilibrage de classes en premier rang moyen, et rééquilibré par la technique SMOTE en second rang). Les résultats du test de Nemenyi montrent que (p > 5%), il n'existe aucune différence entre la performance de deux modèles. La différence apparente dans le classement peut être due à la variabilité aléatoire des données. En ce qui concerne les quatre pipelines restants, les résultats du test de Nemenyi confirment statistiquement la différence de classement avec (*p-value* < 5%).

Conclusion sur le test de Friedman et Nemenyi

Les deux pipelines du modèle de régression logistique utilisant la technique de rééquilibrage (ROSE et SMOTE) dominent la performance de six pipelines testés pour la métrique primaire *F2-Score*, et la métrique secondaire *AUC*. En ce qui concerne la métrique d'évaluation *coût attendu*, les deux

pipelines du modèle Xgboost (sans rééquilibrage de classes, et avec rééquilibrage de classes par SMOTE) sont classés parmi les plus performants selon le rang moyen obtenu.

Le rééquilibrage de classes a un impact sur la performance moyenne du modèle de régression logistique selon les métriques d'évaluation (F2-Score, et AUC). Pourtant l'absence de rééquilibrage de classes des données d'entrainement (Fraude, et Absence de fraude) n'affecte pas le classement de performance moyenne du modèle Xgboost selon la métrique coût attendu.

d) Test de Delong

Les résultats de comparaison de deux aires sous la courbe (AUC) par le test statistique de DeLong sur les mêmes jeux de données montrent que :

- Z = -2,205 (RL_rose, Xgboost_none)

Ce résultat signifie que la différence de AUC est significative entre les deux modèles de classification de fraude (Regression Logistique_rose, XGBoost_none), signe négatif indique la direction [Régression logistique (ROSE) < XGBoost (sans rééquilibrage de classes)].

- (*p : 0,02745*), l'hypothèse nulle H_0 est rejetée. Le modèle Xgboost sans rééquilibrage de classes (AUC 95%), est significativement meilleur pour détecter les fraudes des CSB, que le pipeline du modèle de régression logistique avec équilibrage de classes par la technique ROSE (AUC 85%).

e) Comparaison de métriques de performance de deux pipelines plus performants

Les résultats de principales métriques de performance analysés *(Accuracy, Balanced Accuracy, Kappa de Cohen, NPV, PPV, Sensibilité Spécificité, F2-Score, AUC, Coût attendu)* montrent que le pipeline utilisant le modèle Xgboost sans rééquilibrage de classes présente des performances globalement supérieures au pipeline du modèle de régression logistique avec rééquilibrage de classes par l'algorithme ROSE. Le Xgboost sans rééquilibrage de classes domine les 8 métriques de performance sur 10 analysées. Sur les deux métriques où le pipeline du modèle de régression logistique (ROSE) est plus performant, les résultats montrent que l'écart est négligeable pour la métrique NPV (100% vs 98%), pourtant pour la sensibilité l'écart est de 14%, soit (100% vs 86%).

f) Discussion sur la performance du XGBoost sans rééquilibrage de classes

L'objectif d'utilisation de la détection de fraude, par l'intelligence artificielle est de limiter le nombre de CSB à auditer, dans un cadre décisionnel sensible aux coûts des erreurs du financement basé sur la performance. Dans la majorité des cas, la contrainte de la détection de fraude utilisant l'apprentissage automatique est le déséquilibre de classes. Les résultats de notre étude montrent que la prévalence de fraude des CSB varie de [0% - 20%], dont la prévalence moyenne des 11 enquêtes

réalisées est de 14% pour les CSB du District sanitaire d'Ambatolampy. Les résultats du test de Nemenyi sur les deux pipelines (Xgboost_none vs Xgboost_smote) montrent qu'il n'y a pas de différence statistiquement significative (*p-value=0,13654*), sur l'utilisation de technique de rééquilibrage de classes utilisant l'algorithme SMOTE. En plus le pipeline utilisant le modèle Xgboost avec rééquilibrage de classes par l'algorithme ROSE est classé sixième du classement moyen de performance. Cette différence de classement moyen entre (Xgboost_none & Xgboost_smote) vs (Xgboost_rose) est statistiquement significative selon les résultats du test statistique de Nemenyi ci-dessous :

- Xgboost_rose vs Xgboost_none (p : 2,6 10^{-07}) ;
- et Xgboost_rose vs Xgboost_smote (p : 0,03566).

Performance globale

❖ **AUC**

Il n'existe pas de norme scientifique sur la performance de classification binaire, mais il a y un consensus académique largement reconnu dans la littérature qu'une performance de l'AUC de [90%, 100%] est considérée comme une performance excellente (AUC Xgboost_none : 95%). Il présente une excellente discrimination de cas de fraude vs non fraude » du modèle étudié. Les auteurs comme (A. M. Carrington et al, 2022) [9], et (Fawcett, 2006) [10] ont partagé ce consensus académique pour confirmer la pertinence du modèle pour des applications décisionnelles.

❖ **Exactitude ou (Accuracy)**

L'exactitude (94%) traduit une très bonne performance globale du pipeline. Une exactitude à 94 % indique une bonne capacité du pipeline utilisant le modèle Xgboost sans rééquilibrage de classes à discriminer correctement les classes majoritaire et minoritaire.

❖ **Exactitude équilibrée ou (Balanced Accuracy)**

L'Exactitude équilibrée (91%) indique statistiquement que le pipeline du modèle Xgboost sans rééquilibrage de classes identifie correctement la grande majorité des CSB frauduleux, et les CSB conformes. Le Xgboost (none ou sans rééquilibrage de classes) reste très performant et robuste même en présence de données déséquilibrées, et sans rééquilibrage des données d'entrainement. Ce niveau de capacité élevée de discrimination montre qu'il ne favorise pas la classe majoritaire (CSB conformes). Selon l'analyse proposée par (Brodersen et al, 2010) [11], ce résultat montre qu'une exactitude équilibrée élevée ou proche de 1, traduit une probabilité moyenne élevée de classification de deux classes.

❖ **Coefficient Kappa de Cohen**

Le coefficient Kappa de Cohen à 78%, indiquant un accord substantiel de classification. Il présente une bonne fiabilité de classification. Selon l'échelle proposée par (Landis et al ,1977) [13], le coefficient Kappa de 78% pressente un accord substantiel réel qui confirme la fiabilité statistique du modèle étudié au-delà du hasard.

❖ **Sensibilité et Spécificité**

Le modèle Xgboost sans rééquilibrage de classes présente une bonne capacité de détection de fraudes (sensibilité : 86%), et une précision acceptable (PPV : 76%) indiquant que la majorité des alertes de fraude correspondent à de véritables cas de CSB fraudeurs.

La spécificité est très forte (96%), et permet d'éviter les faux positifs. Et la valeur de NPV (98%) montre une très forte fiabilité de classer les vrais négatifs. Le Xgboost sans rééquilibrage de classes présente une excellente capacité à reconnaître les CSB conformes. Il est efficace pour la détection des fraudes tout en limitant les fausses alertes (spécificité : 96 %, NPV : 98%). La spécificité supérieure à la sensibilité indique qu'il est plus efficace pour identifier les CSB conformes que pour détecter les CSB fraudeurs. Cette situation est fréquemment observée dans les jeux de données fortement déséquilibrés, comme le soulignent (He et at ,2009) [13].

❖ **F2-Score**

Le F2-score est une mesure d'évaluation des modèles de classification qui combine la précision (PPV), et la sensibilité. La valeur de F2-score (84%) est classée comme ayant un très bon niveau de performance, avec un niveau de précision satisfaisant (76%). Le niveau élevé de F2-Score indique que la sensibilité est élevée, et le pipeline Xgboost sans rééquilibrage de classes est efficace pour identifier les comportements frauduleux des Centres de santé de base sous contrat de performance. Cette capacité de détection des CSB frauduleux (F2-Score élevé) est très importante, dans un cadre de décision sensible aux coûts des erreurs, comme le Financement basé sur la performance du secteur sanitaire, où le coût d'une fraude non détectée est plus élevé par rapport au coût d'une fausse alerte**.** Selon Sasaki, 2007[14], le F2-score est une métrique particulièrement pertinente lorsque les couts des faux négatifs sont importants. Rater un cas de fraude est plus grave que produire une fausse alerte.

g) Validation sur des données externes du modèle XGBOOST

La validation a été réalisée sur les données d'achat de performance des CSB du premier trimestre 2024 (T1/2024) du District d'Antanifotsy, avec 22 CSB enroulés dans le mécanisme du FBP.

- Les deux cas confirmés de fraude par l'enquête auprès des usagers ont été détectés par XGBoost. Ces 2 cas confirmés ont ajouté des usagers fictifs dans leurs registres de collecte des données.

L'enquête auprès des usagers a détecté 5 cas de suspicion de fraude dont les 3 cas sont détectés par le modèle XGBoost. Les cas suspects de fraude identifiés par l'enquête doivent être interprétés avec prudence, car les résultats sont fortement influencés par le biais de souvenir, et le biais lié au faible niveau d'instruction des usagers en particulier en milieu rural. Les résultats d'enquête permanente auprès des ménages (EPM) en (2021–2022), (INSTAT, 2024) [15] confirment que 21% de la population n'a aucun niveau d'instruction, et 44,3% de la population a un niveau primaire. Pourtant les 3 cas suspects de fraudes détectés par XGBoost signifient que les données de ces trois CSB sont en faveur d'un risque élevé de fraude.

Pour les 5 CSB ciblés par XGBoost : 2 CSB sont confirmés de fraude, et 3 CSB fortement suspects de fraude. Le modèle XGBoost présente une performance de l'AUC (94.4%).

VIII. Conclusion

Les résultats d'enquête trimestrielle auprès de usagers nous montrent que la prévalence de fraude des CSB ont varié de 0% à 20%, avec une moyenne de 14% pour la période (2019 à 2021). Dans le mécanisme de verification classique ou traditionnelle du FBP, plus de 80% des CSB audités sont conformes, ou sans risque de fraudes. Afin d'optimiser l'efficience de l'approche FBP, cette étude a proposé l'intelligence artificielle pour réduire de 25% par District le nombre de CSB à vérifier. L'IA permet de minimiser les coûts de la verification en ciblant uniquement les CSB susceptibles de présenter de risques élevés de fraudes. L'anomalie des données sanitaires des CSB sous contrat de performance est un indice de risque de fraude.

Après évaluation de performance par une série de tests statistiques (Friedman, Nemenyi, Delong), le modèle de XGBoost est le mieux adapté au contexte FBP à Madagascar. Ce modèle présente une performance globalement élevée en particulier : accurancy (96%), le balanced accuracy (91%), la sensibilité (86%), la spécificité (96%), le F2-score (84%), l'AUC (95%), le coût attendu (36), et enfin le Kappa de Cohen (78%).

La validation externe du modèle avec les données indépendantes de 22 CSB du District sanitaire d'Antanifotsy montre que les cinq CSB classés frauduleux, ou fortement suspects de fraudes par l'enquête auprès des usagers sont identifiés par le modèle XGBoost.

Cette étude montre qu'on peut réduire à 25% de nombre de CSB à vérifier grâce à l'utilisation de l'IA, pourtant une étude économique est nécessaire afin d'évaluer l'impact de l'intelligence artificielle sur la verification ciblée.

L'intégration de tous les résultats d'enquêtes de tous les CSB enroulés dans le mécanisme du FBP comme source de données d'entrainement, et l'amélioration de la qualité des données du système national d'information sanitaire sont indispensables pour mettre en place un système d'IA performant au niveau national.

IX. RÉFÉRENCES

[1] I. Mathauer, E. Dale, M. Jowett, J. Kutzin. (2019). *Purchasing health services for universal health coverage: how to make purchasing more strategic?,World Health Organization*, WHO/UHC/HGF/PolicyBrief/19.6.2019.

[2] CORDAID-SINA, (2010). *PBF in Action : Theories and Instruments, PBF Course Guide, The Hague*. https://www.sina-health.com

[3] C. Achir, A.Douari. (2024). *Le management du risque à l'ère de l'émergence de l'intelligence artificielle, Revue Française d'Économie et de Gestion,* Vol.5: N°1, p.52–77.

[4] P. K. Kamuangu. (2024). (2024). *A Review on Financial Fraud Detection using AI and Machine Learning, Journal of Economics, Finance and Accounting Studies*, 6(1), 67-77.ISSN: 2709-0809.

[5] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander. (2000). *LOF : Identifying density-based local outliers, Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, p. 93–104. doi: 10.1145/342009.335388

[6] G. James, D. Witten, T. Hastie, and R. Tibshirani. (2013). *An Introduction to Statistical Learning : with Applications in R, New York : Springer*, p. 130–136, Chap. 4. ISBN : 978-1-4614-7137-0.

[7] I. Goodfellow, Y. Bengio, A. Courville. (2016). Deep Learning, Cambridge, MA: MIT Press, p. 190–192.

[8] T. Chen and C. Guestrin. (2016). *XGBoost : A scalable tree boosting system, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), San Francisco, CA, USA*,, p. 785–794, . doi: 10.1145/2939672.2939785.

[9] A. M. Carrington, D. G. Manuel, P. W. Fieguth et al., "Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 1, pp. 329–341, Jan. 2022, doi:10.1109

[10] T. Fawcett, (2006). *An introduction to ROC analysis, Pattern Recognition Letters*, 2006.

[11] K. H. Brodersen, C. S. Ong, K. E. Stephan and J. M. Buhmann. (2010). *The balanced accuracy and its posterior distribution, in Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, p.3121–3124.

[12] J. R. Landis and G. G. Koch. (1977). *The measurement of observer agreement for categorical data, Biometrics*, vol. 33, N°. 1, p. 159–174.

[13] H. He and E. A. Garcia. (2009). *Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, N°. 9, p. 1263–1284. doi: 10.1109/TKDE.2008.239.

[14] Y. Sasaki. (2007). *The truth of the F-measure, Teach Tutor Materials*, Vol. 1, N°. 5, p. 1–10.

[15] Institut National de la Statistique. (2024). *Enquête Permanente auprès des Ménages (EPM) 2021–2022, Rapport INSTAT*.